Network Structure Revealed by Short Cycles

James Bagrow,[1, *] Erik Bollt,[2, 1, †] and Luciano da F. Costa[3, ‡]

[1]*Department of Physics, Clarkson University, Potsdam, NY 13699-5820, USA.*
[2]*Department of Mathematics and Computer Science,*
Clarkson University, Potsdam, NY 13699-5815, USA.
[3]*Instituto de Física de São Carlos. Universidade de São Paulo,*
São Carlos, SP, PO Box 369, 13560-970, Brazil

(Dated: February 6, 2008)

This article explores the relationship between communities and short cycles in complex networks, based on the fact that nodes more densely connected amongst one another are more likely to be linked through short cycles. By identifying combinations of 3-, 4- and 5-edge-cycles, a subnetwork is obtained which contains only those nodes and links belonging to such cycles, which can then be used to highlight community structure. Examples are shown using a theoretical model (Sznajd networks) and a real-world network (NCAA football).

I. INTRODUCTION

Complex networks have attracted growing attention because of their non-uniform connectivity patterns, which may give rise to node degree power laws and hubs, known to play an important role in defining several topological properties of the networks [1, 2, 3]. More recently, the fact that many complex networks include *communities*, i.e. sets of nodes which connect more intensely amongst one another than with the rest of the network, has become the focus of increasing attention (e.g. [4, 5, 6, 7, 8, 9, 10, 11, 12, 13]). Indeed, because of statistical fluctuations, even random networks [14, 15] can be found to exhibit communities [16, 17]. Although we still lack a clear-cut definition of a community, the problem of identifying communities in complex networks continues to motivate interest from researchers because of the importance that those structures have for better understanding the general organization of such complex structures (e.g. [18]).

Another important feature of complex networks are the cycles of different lengths which underlie the connectivity of the several models of networks [19]. Actually, the statistical distribution of cycles has been acknowledged as particularly important for defining not only the topology of the respective networks, but also the dynamics of systems running on such frameworks(e.g. [20]). The latter is a direct consequence of the fact that cycles, through feedback, form the scaffolding of memory in dynamical systems.

Generally, the density of cycles tends to increase as more edges are incorporated into a network, with longer cycles being observed earlier than shorter ones (e.g. [21]). Therefore, the density of cycles of different lengths can be used as an indicator of the connectivity between any subset of nodes. In other words, the larger the number of shortest cycles among a subset of nodes, the more connected such nodes are to one another. Longer cycles tend to grow, "coiled up", alongside these shorter cycles, however, blurring the distinction between nodes based solely on short-cycle participation. We present methods to overcome this.

The article starts by presenting the cycle finding algorithm and its application as the core of the community finding algorithm and proceeds by illustrating the application of such a methodology to community finding in a theoretical complex network model (i.e. Sznajd networks [22]) and a real-world football network.

II. DESCRIBING SHORT CYCLES

For a graph $G = \{V, E\}$, $n = |V|, m = |E|$, we are interested in finding cycles of length 3,4, or 5 containing some starting vertex $v \in V$. To describe these cycles we begin by decomposing G into shells S_i about v. We define shell S_i to be the set of all vertices (and edges between those vertices) at a distance i from the starting vertex v. Since we are only interested in cycles of length ≤ 5, we need only to keep shells S_1 and S_2.

It is simple to describe all possible short cycles using these shell decompositions. For example, for every edge e_{ij} in S_1 about v, there exists a 3-cycle (triangle) v–i–j–v. Similarly, for every path of length 2 or 3 in S_1, there exists a 4- or 5-cycle, respectively. Another 4-cycle and two more 5-cycles exist involving both S_1 and S_2.

In general, for a cycle of length $L \geq 3$, the number of such possible "cases" grows with L. Since it requires 2 edges to visit a shell, an L-cycle can visit at most J shells, where

$$J = \begin{cases} \frac{L}{2}, & L \text{ even}, \\ \\ \frac{L-1}{2}, & L \text{ odd}. \end{cases} \tag{1}$$

If the farthest shell the cycle visits is S_j, $j < J$, there are at most $L - 2j$ remaining edges that must be distributed between and within the $S_1, S_2, ...S_j$ shells. The

[*] Electronic address: bagrowjp@clarkson.edu
[†] Electronic address: bolltem@clarkson.edu
[‡] Electronic address: luciano@if.sc.usp.br

number of ways to distribute $L-2j$ edges over j shells is $\frac{(L-2j+j-1)!}{(L-2j)!(j-1)!}$. However, it is possible for a cycle to "zig-zag" between shells, using more than the $2j$ edges necessary to visit the j shells. Therefore, the total number of possible ways to distribute an L-cycle is at least:

$$N_l(L) = 1 + \sum_{j=2}^{J} \sum_{i=0}^{J-j} \frac{(i+j-2)!}{i!(j-2)!} \frac{(L-2i-j-1)!}{(L-2j-2i)!(j-1)!}, \qquad (2)$$

with the outer sum accounting for all the possible shells the cycle can visit, the inner sum for all the optional pairs of edges that can lie between shells and the +1 for the one possible cycle that visits the first shell only. Here, i is the number of pairs of edges between shells beyond the j necessary to visit the j shells.

This calculation fails to take into account permutations of the *ordering* of edges between and within two adjacent shells. A simple upper bound is possible, however, as there are certainly no more than $L!$ possible permutations over the whole network:

$$N_u(L) = 1 + \sum_{j=2}^{J} \sum_{i=0}^{J-j} \frac{(i+j-2)!}{i!(j-2)!} \frac{(L-2i-j-1)!}{(L-2j-2i)!(j-1)!} L!, \qquad (3)$$

with

$$N_l(L) \le N(L) \le N_u(L). \qquad (4)$$

III. CYCLES AND COMMUNITIES

For a graph G, a cycle C is a subset of the set of edges E containing a continuous path, where the first and last node of the path are the same [23]. Permutations of cycles may be ignored since we will be working exclusively with sets of edges. Throughout this work, we limit ourselves to short cycles, typically those of length l, $3 \le l < 6$. These shorter cycles may provide the advantage of faster calculation times.

Community structure can be studied by comparing the edges covered by these cycles with the original graph. Let

$$C_l(i) \equiv \text{the set of edges traversed by all } l\text{-cycles starting from vertex } i \qquad (5)$$

Starting from all vertices and limiting ourselves to only short j-cycles [29],

$$C \equiv \bigcup_{i \in V} \bigcup_{j} C_j(i). \qquad (6)$$

Then, for a graph G, we construct a graph H where,

$$H = \{V, E \setminus C\} \qquad (7)$$

is the graph G containing only edges that do not participate in j-cycles. Separate communities in G will appear as disconnected components in H. We interpret vertices with degree zero in H as communities of size one.

In specifying H, the question of what to choose for j has been left open. For example, choosing just $j = \{3\}$ will correspond to deleting all edges from G that participate in 3-cycles, generally not a useful result. One may consider j to be a tunable parameter, used to get a desired result when applied to a specific network.

One issue that can occur is that longer cycles often overlap shorter cycles. In terms of communities, most inter-community edges contain few (if any) short cycles, but intra-community edges tend to contain both long and short cycles, since a long cycle can "coil" inside the community. If one were to just delete all 5-cycles in a graph, it is very possible to end up deleting all edges.

There is quite a bit of leeway in how we choose j and build H, and we can use this to our advantage. For example, pick two cycle lengths s and t, $s < t$ and compute C_s and C_t. Then, build another set of edges, $C_{t\setminus s}$

$$C_{t\setminus s} \equiv C_t \setminus C_s, \qquad (8)$$

containing the set of edges that participate in t-cycles *but not* s-cycles. The graph $H = \{V, C_{t\setminus s}\}$ will contain edges that tend to be between communities and not within, for an appropriate choice of t and s. One can think of this as a "backbone" of the network, and deleting these edges may be a useful pre-processing step for applying other community-detection algorithms, including betweenness [4, 10].

IV. APPLICATION EXAMPLES

We present example applications of the methods presented in Section III to two networks: a network of NCAA Division I-A football games held during the 2005 regular season [30] and a Sznajd network [24]. In addition, we discuss how these methods can break down and ways to overcome that.

A. Football Network

In NCAA football, teams are grouped into *conferences* based on location. To save on transportation time and cost, more games are played between teams in the same conference than in different conferences. Thus, a graph of the game schedule, where nodes are teams and edges connect teams that have played against each other, naturally exhibits community structure based on these conferences [25].

Figure 1a displays the original network, call it G. As a first pass, let's use $j = \{3\}$ and generate $G_3 = \{V, C\}$, pictured in Figure 1b using the same layout as 1a. This deletes all edges that do not participate in 3-cycles.

Most deleted edges are between conferences, though some edges remain. This will not split the network into seperate components based on the communities but it may be useful as a preprocessing step for betweenness or another community detection algorithm.

In addition, let us build $C_{t\backslash s}$, as per Equation 8. For this network, we have chosen $t = 5, s = 3$. Figure 1c shows $G_{5\backslash 3} = \{V, C_{t\backslash s}\}$, again using the same layout as 1a. For improved clarity, Figure 1d shows $G_{5\backslash 3}$ with a layout emphasizing that all edges are between conferences.

We propose that edges in $C_{5\backslash 3}$ comprise the majority of this network's inter-community structure. To test this, one can compare the distributions of edge betweenness for these backbone and non-backbone edges, as shown in Figure 2a. Backbone edges tend to carry much higher betweenness values than the more common non-backbone edges.

B. Sznajd Network

One particularly interesting category of complex networks are the so-called *geographical models* (e.g. [27, 28]), whose nodes have well-defined positions in an embedding metric space S. Typically, the connectivity in such networks is affected by the adjacency and/or the distance between pairs of nodes, with nodes which are closer one another having higher probability of being connected. As an immediate consequence of such an organizing principle, communities in traditional geographical communites are closely related to the presence of spatial clusters of nodes, i.e. groups of nodes which are closer one another than with the rest of the network. Introduced recently, the family of geographical networks known as *Sznajd networks* [22] allow rich community structure as a consequence of running the Sznajd opinion formation dynamics [24] among the network edges instead of considering the states associated to each network node. Starting with a traditional geographical network (called the *underlying network* Γ) where the connections are defined with probability proportional to the distances between pairs of nodes, a percentage of edges of Γ are removed, yielding the initial condition for the Sznajd dynamics. Then, edges from Γ are chosen randomly and used to influence the respective surrounding connectivity. For instance, in case the chosen edge (i, j) is on (i.e. it does correspond to a link in the current growth stage), the edges in Γ which are connected to the nodes i and j are established with probability p. An analogue procedure is considered with respect to edges which are absent. In order to avoid convergence to the trivial ground states where all edges are set on or off, the dynamics also consider as feedback the total number of established edges.

Figure 3a shows a Sznajd Network. Edges that do not participate in 3-cycles are indicated. As can be seen, many of these edges fall "outside" of the more dense regions of the network. This is a good first pass, and may be used to initialize another algorithm, similar to our football result, but it will not give detailed information on the hierarchical community structure.

Figure 3b shows the same network as 3a, but with the edges of $C_{5\backslash 3}$ highlighted. One can imagine removing both the C_3 and $C_{5\backslash 3}$ edges to further enhance the separation.

V. CONCLUDING REMARKS

The identification and characterization of the communities present in complex networks stands out as one of the most important approaches for understanding their structure and possible formation and evolution. At the same time, the distribution of cycles of various lengths in a complex network has important implications for the connectivity, resilience and dynamics of the respectively studied networks. The current work brought together these two important trends, in the sense of applying short cycle detection as the means to help the identification of communities in complex networks. The suggested methodology has been applied with promising results to the identification of communities in a theoretical network model, more specifically a Sznajd geometrical networks, as well as to a real-world network (NCAA).

The relationship between the cycles and communities in the football network has been further investigated in terms of the betweeness centrality measurement, confirming that the obtained backbone edges tend to exhibit higher betweeness values.

Acknowledgments: L. da F. Costa thanks FAPESP (05/00587-5) and CNPq (308231/03-1) for financial support.

[1] S.-H. Yook, H. Jeong, and A.-L. Barabási, Proc Natl Acad Sci USA **99**, 13382 (2002).

[2] S. N. Dorogovtsev and J. F. F. Mendes, Advances in Physics **51**, 1079 (2002), cond-mat/0106144.

[3] S. Bornholdt and H. G. Schuster, eds., *Handbook of Graphs and Networks: From the Genome to the Internet* (John Wiley & Sons, Inc., New York, NY, USA, 2003), ISBN 3527403361.

[4] M. Girvan and M. E. J. Newman, Proc Natl Acad Sci USA **99**, 7821 (2002).

[5] M. E. J. Newman and M. Girvan, in *Statistical Mechanics of Complex Networks*, edited by R. Pastor-Satorras, J. Rubi, and A. Diaz-Guilera (Springer, Berlin, 2003).

[6] M. E. J. Newman, The European Physical Journal B **38**,

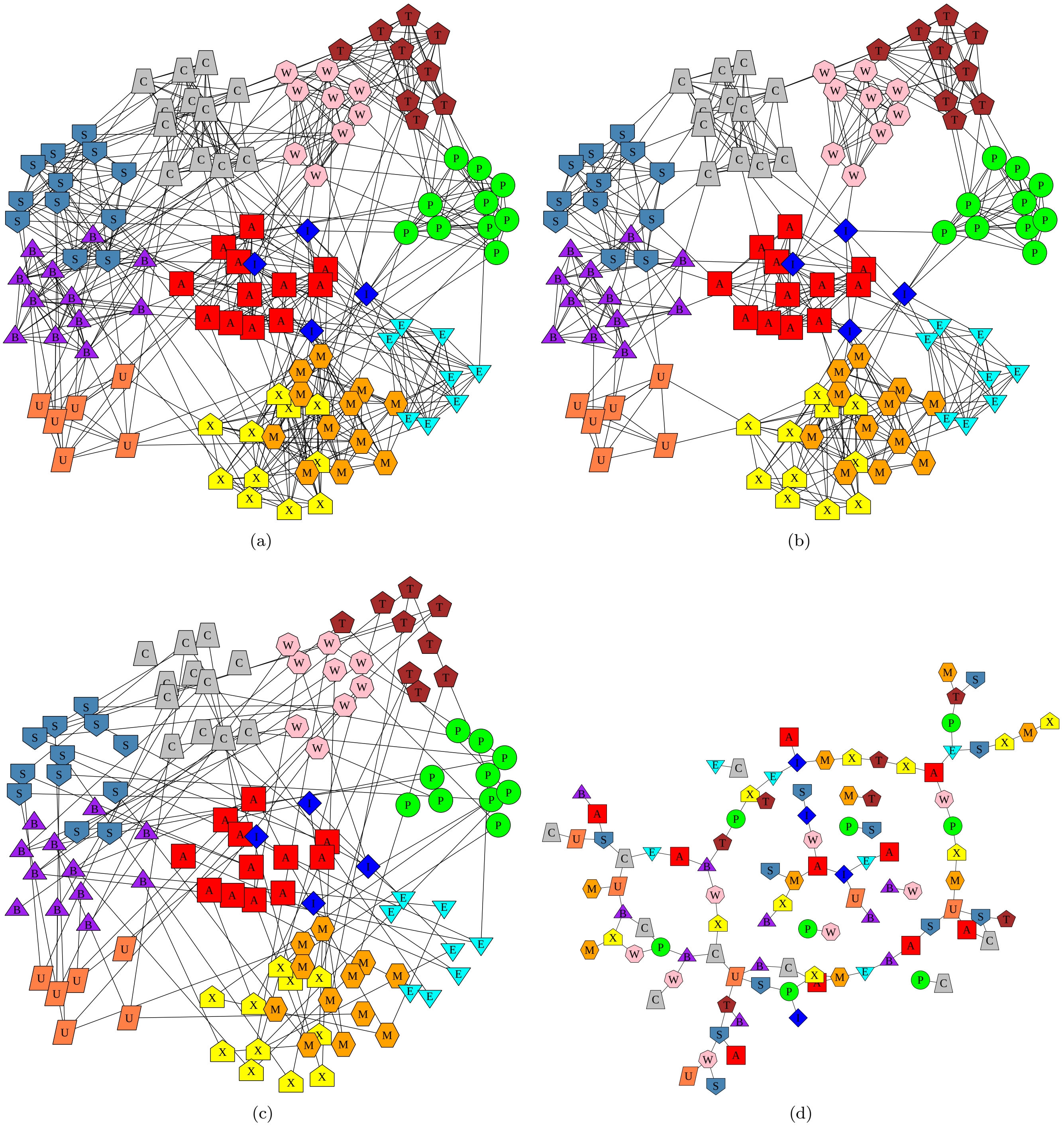


FIG. 1: (color online) The NCAA Div I-A 2005 regular season with all edges (a), with 3-cycles only (b), and with just $C_{5\backslash 3}$ edges (c). (d) is the same graph as (c) but with a layout emphasizing that no edges within conferences remain (degree zero nodes omitted). As per [26], the conferences are: A = Atlantic Coast, B = Big 12, C = Conference USA, E = Big East, I = Independent, M = Mid-American, P = Pacific Ten, S = Southeastern, T = Western Athletic, U = Sun Belt, W = Mountain West, X = Big Ten.

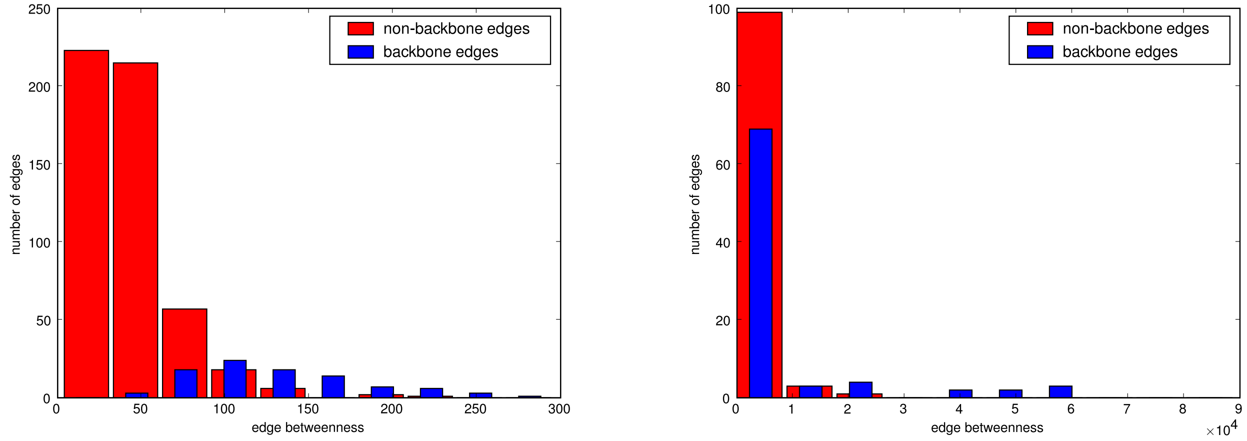
321 (2004).

[7] M. E. J. Newman and M. Girvan, Physical Review E **69**, 026113 (2004).

[8] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).

[9] A. Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E **70**, 066111 (2004).

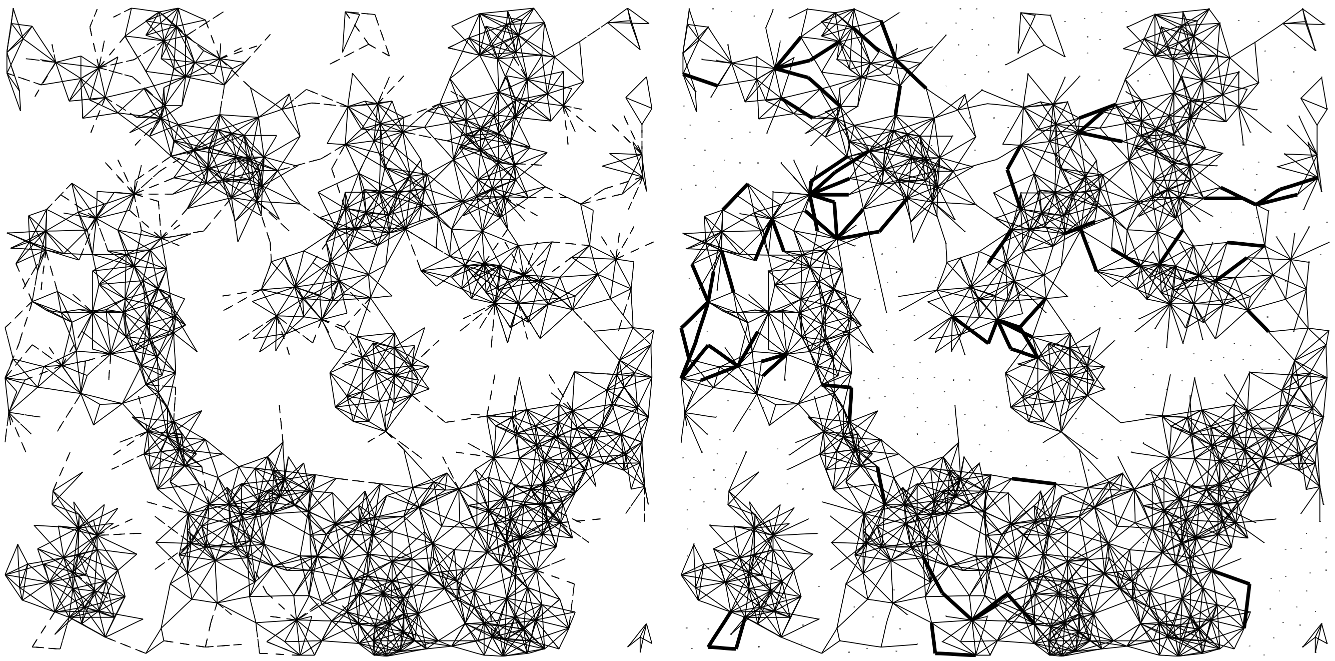
[10] J. P. Bagrow and E. M. Bollt, Phys. Rev. E **72**, 046108 (2005), cond-mat/0412482.



(a)

(b)

FIG. 2: (color online) Histogram of edge betweenness for non-backbone edges (red) and backbone edges (blue) for the NCAA 2005 football network (a) and the Sznajd network shown in Figure 3 (b). For the football network, the mean (unnormalized) betweenness is 42.8 for non-backbone edges and 132.9 for backbone edges. Note that backbone and non-backbone histograms use the same bins; the front-most bins have been narrowed for clarity. The Sznajd non-backbone bins have also been scaled down by a factor of 25 for clarity.



(a)

(b)

FIG. 3: A Sznajd network. Edges that do not participate in 3-cycles are dashed (a). Edges in $C_{5\backslash 3}$ are bold (b). Note that nodes of degree zero have been omitted for clarity.

[11] M. E. J. Newman, Proc Natl Acad Sci USA **103**, 8577 (2006).

[12] M. E. J. Newman, Phys. Rev. E **74**, 036104 (2006).

[13] M. A. Porter, P. J. Mucha, M. E. J. Newman, and A. J. Friend, submitted to Social Networks (2006), physics/0602033.

[14] P. Erdös and A. Rényi, Publ. Math. **6**, 290 (1959).

[15] B. Bollobás, *Random Graphs* (Academic Press, London, 1985).

[16] R. Guimera, M. Sales-Pardo, and L. A. N. Amaral, Phys. Rev. E **70**, 025101 (2004).

[17] J. Reichardt and S. Bornholdt (2006), cond-

mat/0606295.

[18] R. Guimera and L. A. N. Amaral, Nature **433**, 895 (2005).

[19] H. D. Rozenfeld, J. E. Kirk, E. M. Bollt, and D. ben Avraham, J. Phys. A **38**, 4589 (2005), cond-mat/0403536.

[20] A. Arenas, A. Diaz-Guilera, and C. J. Perez-Vicente, Physical Review Letters **96**, 114102 (2006), cond-mat/0511730.

[21] L. da Fontoura Costa, Physical Review E **70**, 056106 (2004), cond-mat/0312712.

[22] L. da F. Costa, Intl. J. Mod. Phys. C **16**, 1001 (2005).

[23] B. Bollobás, *Modern Graph Theory* (Springer, New York, 1998).

[24] K. Sznajd-Weron and J. Sznajd, Intl. J. Mod. Phys. C **11**, 1a57 (2000).

[25] T. Callaghan, M. Porter, and P. Mucha, accepted in American Mathematical Monthly (2003), physics/0310148.

[26] J. Park and M. E. J. Newman, J. Stat. Mech. **P10014** (2005), physics/0505169.

[27] M. T. Gastner and M. E. J. Newman, The European Physical Journal B **49**, 247 (2006).

[28] L. da F. Costa and L. A. Diambra, Physical Review E **71**, 021901 (2005).

[29] Indeed, here we specify short cycles as those of length 3, 4, or 5 but this is not a set rule and, in certain circumstances, it may prove advantageous to consider 4- or 5-cycles, or even just 5-cycles.

[30] Data taken from published schedule at `http://www.ncaa.org`